Quality Assurance Project Plan (QAPP)

The Determination of Cumulative Risk to Phthalates from Biomonitoring Data

Prepared by

Paul Price National Exposure Research Laboratory (NERL)

Office of Research and Development (ORD)

U.S. Environmental Protection Agency (EPA)

Research Triangle Park, NC 27711

June 10, 2016

The organization implementing this project is the U.S. Environmental Protection Agency (EPA)

Approved by:

PAUL PRICE Digitally signed by PAUL PRICE DN: c=US, o=U.S. Government, ou=USEPA, ou=Staff, cn=PAUL PRICE, dnQualifier=0000075735 Date: 2017.06.09 11:14:05 -04'00' Paul Price Date Principal Investigator	JEANETTE REYES (affiliate) Digitally signed by JEANETTE REYES (affiliate) Date: 2017.06.12 08:12:52 -04'00' Jeanette Reyes Date Researcher
Andrew Gillespie Date CED Division Director	James Noel Date QA Manager

Contents

Part A. Project Management

A1. Description of the EPA Quality System

The EPA requires that all data collected for the characterization of environmental processes and conditions are of the appropriate type and quality for their intended use. This is accomplished through an Agency-wide quality system for environmental data. Components of the EPA quality system can be found at http://www.epa.gov/quality. EPA policy is based on the national consensus standard ANSI/ASQ E4-2004 Quality Systems for Environmental Data and Technology Programs: Requirements with Guidance for Use.

This QAPP provides information concerning the development and application of the Maximum Cumulative Ratio metric as a tool to investigate the contribution of individual phthalates to cumulative risks posed by six phthalates as measured by ten years of biomonitoring.

A2. Distribution List

Paul Price, EPA/ORD/NERL/CED/IO

Jeanette Reyes, EPA/ORD/NCEA/EMAG

James Noel, EPA/ORD/NERL/CED

A3. Project/Task Organization and Responsibilities

This Quality Assurance Project Plan (QAPP) was developed with reference to Guidance for Quality Assurance Project Plans for Modeling EPA QA/G-5M (U.S. Environmental Protection Agency, 2002).

The project is an analysis of cumulative risks based on biomonitoring data for six phthalates using the Maximum Cumulative Ratio. The project is in support of research on cumulative exposure under Section 6.234 of the Human Health Risk Assessment program of EPA. The following list provides the project participants along with their respective responsibilities.

Dr. Paul Price (EPA) is the principal investigator (PI) in the project who is responsible for designing the overall research project to assess risks from cumulative exposures.

Dr. Jeanette Reyes (Oak Ridge Institute for Science and Education (ORISE) Research Participation Program, hosted at U.S. Environmental Protection Agency, Research Triangle Park) is responsible for conducting the technical aspects of the research projects, including, reviewing literature, extracting data from published sources and databases, performing the analyses that generate the outputs of exposure, risk, and risk related metrics, analyzing model results, and writing/communicating research design and findings.

James Noel (EPA) is the EPA NERL CED Quality Assurance Managers who are responsible for reviewing and approving this QAAP as well as providing any project assessments on an as needed basis.

A4. Problem Background

Exposure and risk assessors use algebraic models to estimate metrics of exposure and risk using existing data from multiple sources. The creation of these metrics involve the use of mathematical models (algebraic equations) that have been previously established in peer reviewed publications and or governmental guidance documents. The metrics, and the data used in their creation, are analyzed using statistical tests, and are sorted and visualized using a range of plotting techniques. The results are published in peer reviewed journals. The data can take the form of large and complex data sets such as the National Health and Nutrition Examination Survey (NHANES) data on biomonitoring (CDC, 2016) or published values of parameters in the models that generate the metrics. In these projects, the new findings arise not from the creation of the data, the development of new models of the metrics, the development of new statistical tests, nor the development of new tools for data visualization, but rather from the novel application of existing tools to the data set. This project, an analysis of cumulative risks based on biomonitoring data for six phthalates using the Maximum Cumulative Ratio, is such a project.

The project obtains data from existing public sources and performs the following tasks:

- The identification and extraction of data from credible public sources;
- Application of existing algebraic models to the data to produce exposure and risk metrics;
- Perform statistical analyses of the data and metrics using "off the shelf" publically available statistical software; and
- Perform data visualizations using "off the shelf" publically available software.

The scope of the QAPP is limited to the above actions. The quality of the data, the methods of calculating the metrics, and the statistical tools are assumed to be addressed by separate QAPPs.

This QAPP focuses on three questions.

1. Are the data and the proposed models, tests, and visualizations appropriate for the proposed project?
2. Are the data and all calculations clearly described?
3. Are calculations performed as described?

A5. Task Description and Schedule

The project will download biomontoring data from the National Health and Nutrition Examination Survey (NHANES) web page (https://wwwn.cdc.gov/nchs/nhanes/) on the level of urinary metabolites of six phthalates (CDC, 2016). These data are used to estimate daily exposures for six phthalates in the surveyed individuals. The estimations will be performed using published methodologies and published data on the metabolism of the compounds (Christensen et al. 2014). The estimates of daily exposures are then used to perform a screening analysis of risk from combined exposures and an assessment of the relative contribution of individual phthalates to individuals' combined risks using previously published methods (EPA, 2006; Price and Han, 2011; Price et al. 2012). All data manipulations, analysis, and visualizations were conducted in R using version 3.2.2 using the R packages ggplot2 (version 2.2.0) and survey (version 3.31). The results of the analyses will be published in a peer review publication.

A6. Quality Objectives and Criteria

The Quality Objectives of the projects covered by this QAPP are:

1. Demonstrate that the data and the proposed models, tests, and visualizations appropriate for the proposed project;
2. Demonstrate that the data and all calculations clearly described; and,
3. Demonstrate that all calculations are performed as described.

The first Quality Objective is addressed by the development of clear and logical arguments for the findings of the research project that take into account the strengths and limitations of the models, statistical tests, and data visualization tools. This objective is achieved by following the criteria and guidance established in Section D. The criterion for satisfying this objective is the creation of a final report (i.e. one or more publications in peer reviewed journals) that links the data and calculations to specific findings in a logical and objective fashion.

The second Quality Objective is achieved by the careful documentation of the processes used in the project and is discussed in A8 below. The criterion for success is the ability of an assessor outside of the project being able to duplicate all analyses presented in the final report beginning with the extraction of the raw data.

The final Quality Objective is achieved by independently checking the results to confirm the analyses. The methods used to perform the checks are described in Section B. The criterion for success will be to perform and independent analysis that matches to four decimal places the results of the calculations. If the metric is determined using probabilistic model the criterion for matching results should be based on the stability of the model outputs.

A7. Special Training Requirements/Certification

Specialized expertise and qualifications needed to participate in this project include an advanced level of graduate training (M.S. or Ph.D.) and experience in the fields of mathematics, statistics, engineering, chemistry, physics, or environmental sciences. For trainees working on the project, a basic knowledge of programming and mathematics and the ability to identify and extract relevant data and information from articles and databases will be required.

When a new individual becomes involved in the project, this QAPP will be sent to him or her in order to review the quality control and quality assurance procedures required by the lab. The individual will be encouraged to ask questions should any part of this document be unclear, and upon reading their name(s) and dates of completion will be entered into an electronic record.

A8. Documentation and Records

In an effort to facilitate transparency, reproducibility, and credibility, the research team will store all data sets, codes, interim values, and parameters as electronic records on EPA's computer system. A copy of all information on a project will be stored in a shared directory specific to the project on the EPA L:// drive (\\AA.AD.EPA.GOV\ORD\ORD). This directory will be accessible to the PI and project members working on the individual project. The data stored in this folder and will be retained as indicated by EPA record schedule 1035. The directory will include each of the following folders.

Data Cumulative Phthalate Project. This folder should include the following:

- Original documents (e.g., html files, text files, data files, PDFs) as obtained from the external source.
- The raw data in the form used by subsequent analyses (e.g., .csv files, R data tables, or SAS data files).
- Documentation of the origin of the data. If this source is the internet, then the date of access will be recorded.
- A data dictionary for the raw data.

Literature Cumulative Phthalate Project. This folder will include .PDF files for relevant articles and technical documents (if applicable).

QA Cumulative Phthalate Project. This folder should include electronic copies of the QAPP, documentation of familiarity of project member with the QAPP, and files related to QA checks of the calculations. These files should be sufficiently documented so as to allow other researchers to duplicate the QA check.

Calculations and analyses Cumulative Phthalate Project. This folder includes the documentation of the calculations that use the raw data, interim findings, and results. The data will be stored in standard machine-readable format (typically a comma separated value, CSV, file) that is easily read by different programming languages and can be easily disseminated to others via required EPA systems (e.g., ScienceHub). A data dictionary will be included within a .CSV file that defines the source and units of values for each model parameter.

Model Codes Cumulative Phthalate Project. This folder will contain a copy of the final data model. It will contain all codes and inputs that would be required to reproduce the analyses. The documentation should be saved as either a .TXT or a .PDF file (or can be read as text file). The model code should be documented at a sufficient level of detail so as to allow the PI or other person trained on the project to understand how it is functioning. This may include descriptions of both subroutines and individual lines of code. The definitions of parameters from the data dictionary will be recorded as annotations within the model code.

Software used Cumulative Phthalate Project. This folder will include a document in a .DOC file identifying all software used in the project. The document will specify version numbers of software, specific methods, programs, applications used, and settings or options chosen. Data sets used by the software will be described. Documentation should be sufficient to allow others to duplicate the analyses.

The analyses and graphical representations will be summarized in one or more manuscripts that will be submitted for publication in peer-reviewed journals. The manuscripts will serve as the report that documents the various steps in the project. The manuscripts will be sufficiently detailed that an experienced modeler, reviewer, or interested end users are able to reconstruct the manuscripts' finding from the original sources of the data. The supplemental information for the manuscripts, stored by the journal and/or on the EPA Science Hub, will include the raw data and will be sufficient to allow the recreation of the project's findings. The manuscripts will outline how the data and analyses provide a fit basis for the project's findings.

Part B. Determination of Exposure and Risk Metrics and use of Statistical and Data Visualization Tools

The project as covered by this QAPP begins after the project team has defined the project goals, the sources of data, the desired metrics, and the potential tools for statistical analyses and visualization. This process is likely to have occurred by an iterative process that considered various data sets and tools. This decision-making process often requires the expertise of the investigator and a great deal of professional judgement.

B1. Data Extraction

Existing data can take the form of large empirical data sets that reflect inter-individual variation in exposure, temporal and spatial variation in sources of exposure, and variation across chemicals. Data can also be individual values for model parameters. The Quality Objectives for this step of the process will be satisfied by the documentation requirements and stored in the Data for Project folder according to section A8

B2. Data Manipulation

The data manipulation occurs during the determination of the exposure and risk metrics and in the organization of data in preparation for the use of statistical analyses or data visualizations. The process will be performed using a standard programming language or statistical software. Therefore, data will be coded in commercially available software for which the EPA investigators have valid licenses, e.g., Excel (Microsoft Inc.), SAS (SAS, Inc.), MATLAB (the MathWorks Inc.) or open source environments that are approved for use on EPA machines, e.g. R (https://www.r-project.org/) or Python (https://www.python.org) and will be described in Software Uses of Project as specified in section A8.

The Quality Objectives for data manipulation are to clearly describe the manipulations and to demonstrate that they are performed as described. Clarity in the description is achieved by following the requirements in section A8, in particular the Calculations and Analysis project folder.

The accuracy of data calculations is confirmed by performing a QA check. This can be performed in a number of ways, but typically another member of the team will either independently conducts the same calculations to arrive at the same answer as that using the original data set to four decimal places, or carefully follows both the logic and the math of the original calculations. When the equations are simple, an appropriate evaluation approach would include two independent codings of the equations (even in different languages), and a comparison and reconciliation of the results. The results of the QA Check will be saved in the QA Data project folder.

Quality assurance checks will be included in the model code when possible and appropriate. For example, checks on reasonable bounds for parameters and descriptors will be written into the code, and estimated consumer product weight fractions will be checked for summing to 1. All of these checks will be documented in the model code.

B3. Statistical Analyses and Data Visualization

Once the risk metrics are determined, the results are analyzed by performing statistical tests and procedures (e.g., determination of moments of a distribution, correlation analyses, and comparisons of data). Findings of the analyses are often communicated by graphical representations of the data. Software programs (both commercial and open source) are available to perform the tests and to create the graphics. The relevant Quality Objective for the use of the programs that perform the statistical analyses and data visualization is to clearly describe the use of the software. Clarity in the description is achieved by following the requirements in section A8.

Part C. Assessment and Oversight

C1. Assessment and Response Actions

At the discretion of the researchers as well as the Quality Assurance Manager a mid-project technical system audit may be requested by the project Quality Assurance Manager to assess the quality assurance process of the project. Any findings, research best practices, or improvements will be compiled in a report and filed with other QA documents. Any corrective actions and responses will be documented and filed with the Assessment Report. These findings and corrective measures will be located in the office of the Principal Investigator (PI).

C2. Reports to Management

The PI, and other team members will meet on a regular basis (typically weekly or bi-weekly) to discuss progress of data acquisition, data manipulation, and graphic presentations. A final formal report in the form of one or more publication-ready manuscripts and supplemental materials for uploading to ScienceHub (https://sciencehub.epa.gov/sciencehub/) will be available at the completion of each project (see Task Descriptions in section A5).

Part D. Data Validation and Usability

D1. Analysis Review and Prediction Reliability

Data analyses are iterative in nature. Errors are often encountered when calculating metrics or analyzing data. Attention should be given to consistency between steps of analyses, across related findings, and between graphic and tabular presentations of the same data. In the case of statistical analyses multiple techniques can be applied to a data set that provide independent but overlapping/related information. Consensus from of multiple lines of analysis are preferred of single test results because of the reduced chance of a single error driving a finding.

The PI, and other team members (when applicable) will discuss the status of the various individual analyses performed in a project during regularly-scheduled meetings (section C2), and once it is unanimously decided that the overall analysis can no longer be improved, it will be considered in its final form for documentation in PDF or Word format (i.e. one or more manuscripts). The conclusion that the overall analysis can no longer be improved past its current form will be based on correctness of inputs and parameters into the metric determinations, the availability of any additional descriptor or other data that could improve the analysis, and the likelihood that results from the use of additional statistical and data visualization tool would change the project's findings.

D2. Analysis Evaluation

Published literature, other databases, and data obtained from publically-available online sources data should be referenced and cited in electronic records (see section A6). Data from publically-available sources and outside databases must be evaluated for accuracy and quality and only used for applications consistent with its level of quality, and this evaluation should be documented in the final report (e.g., in peer-reviewed publications).

Christen et al., 2014 reported findings on phthalates doses and risk metrics for a previous, single cycle of biomonitoring data from NHANES. These findings will be duplicated using code developed for this project. By duplicating results from previous publications of other researchers using data we extracted and metrics we independently calculated, we increase the confidence of the code and confirms portions of the analysis.

Part E. References

CDC. 2016. National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, [accessed 3 January 2017]. Christensen KLY, Makris SL, Lorber M. 2014. Generation of hazard indices for cumulative exposure to phthalates for use in cumulative risk assessment. Regul. Toxicol. Pharmacol. 69:380–389; doi:10.1016/j.yrtph.2014.04.019.

EPA. 2007. Concepts, Methods, and Data Sources for Cumulative Health Risk Assessment of Multiple Chemicals, Exposures and Effects: A Resource Document (Final Report).; doi:EPA/600/R-06/013F.

Price P, Han X, Junghans M, Kunz P, Watts C, Leverett D. 2012. An application of a decision tree for assessing effects from exposures to multiple substances to the assessment of human and ecological effects from combined exposures to chemicals observed in surface waters and waste water effluents. Environ. Sci. Eur. 24:34; doi:10.1186/2190-4715-24-34.

Price P and Han X. 2011. Maximum Cumulative Ratio (MCR) as a Tool for Assessing the Value of Performing a Cumulative Risk Assessment Int. J. Environ. Res. Public Health 8, 2212-2225; doi: 10.3390/ijerph8062212 (http://www.mdpi.com/1660-4601/8/6/2212/)